RNA-タンパク質複合体構造予測のための 3D-CNN による評価モデルの開発

谷村直樹[i]，石田純一[ii]，浜田道昭[iii]

Development of an Evaluation Model for RNA-Protein Complex Structure Prediction using 3D-CNN

Naoki TANIMURA, Junichi ISHIDA, Michiaki HAMADA

RNA-タンパク質複合体の構造予測がシミュレーションを用いて行われているが，結果として得られる多数の構造から正解を特定することは未だ容易ではない．この課題の解決に向けたデータ駆動型の手法を提案する．本研究では，複合体の 3 次元構造を学習データとし，3D-CNN（3 次元畳み込みニューラルネットワーク）を用いて構造の妥当性を評価するモデルを構築した．この評価モデルは，既存のシミュレータで提示されるエネルギー指標と同等の性能を有する．さらに，説明可能な AI（XAI）技術を適用することで，評価の根拠を物理化学的に解釈できる可能性も示した．本技術の発展により，創薬などに向けた RNA アプタマー配列の効果的・効率的な設計プロセスの構築に大きく貢献することが期待される．

(**キーワード**): RNA-タンパク質複合体構造予測，ボクセル，3 次元畳み込みニューラルネットワーク，説明可能な AI，デジタルトランスフォーメーション

1 はじめに

RNA-タンパク質複合体は，遺伝子発現制御，RNA スプライシング，翻訳，ウイルス複製など，様々な細胞プロセスにおいて極めて重要な役割を果たしている．これらの構造とその相互作用を正確に予測することは，これらのプロセスの分子メカニズムを理解するために不可欠であり，創薬やバイオテクノロジーにおいて重要な意味を持っている．生体分子構造を特定するための X 線結晶構造解析やクライオ電子顕微鏡などの実験技術は大幅に進歩してきたものの，RNA-タンパク質複合体の 3 次元構造を原子レベルの解像度で決定することは，RNA の構造が揺らぎやすいこと，RNA のサイズが低分子と比べて大きいことから依然として困難となっている．そのため，世界最大のタンパク質構造データベースである Protein Data Bank (RSCB PDB)[1)]においても，RNA-タンパク質複合体の構造データは非常に少ない．

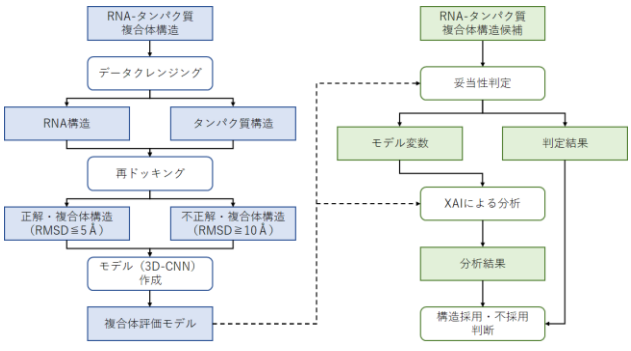

図 1 本報告の概要（左：RNA-タンパク質複合体構造評価モデルの構築，右：評価モデルの利用）

[i] サイエンスソリューション部 デジタルエンジニアリングチーム 次長 博士（理学）
[ii] サイエンスソリューション部 デジタルエンジニアリングチーム コンサルタント 博士（工学），（現）Matlantis 株式会社 テクニカルソリューション部 カスタマーサクセスチームリーダー 博士（工学）
[iii] 早稲田大学 理工学術院 先進理工学部 電気・情報生命工学科 / 先進理工学研究科 電気・情報生命専攻 教授

このような状況のもと，計算により RNA-タンパク質複合体構造の予測と検証を行う手法が研究されてきた．代表的なものとして，ドッキング，分子動力学などのシミュレーション，スコアリング関数などの手法があり，実験で得られた複合体構造データを再現するようなシミュレーション，スコアリング関数が検討され，成果を上げてきた．しかしながら，原子数に比例して増大する自由度，モデルの精緻化と計算コストのトレードオフやスコアリング関数を正確に定義することが困難であることなどが，依然として重要な課題となっている．

これらの課題は，シミュレーションによって得られる多数の複合体構造候補の中から妥当な構造を選択する際の困難に直結している．現状では，多数の複合体構造から不確かな判断基準に従って人手で処理できる程度に候補数を絞り込む必要があり，最終的には 3 次元構造ビューワを用い化学的な知見に基づいて構造を確認し，複合体構造候補の選択が行われている．

近年分子構造モデリングにおいても，データ駆動型の手法として，機械学習アプローチ，特に深層学習が適用されてきており，タンパク質構造[2]，リガンド結合[3]，タンパク質間相互作用[4]の予測などにおいて目覚ましい結果が得られている．また分子構造成立の背後にある物理化学現象の理解・解明については，従来原理・法則に基づき構築されたシミュレーションやスコアリング関数に基づいて行われてきたが，深層学習に対しては説明可能な AI（XAI : Explainable AI）技術の発展がみられており，深層学習モデルの予測の要因を可視化し説明性を高める手段として活用することが可能になってきている．RNA-タンパク質複合体の予測においては教師データとなる構造データが少ないという課題を克服しつつ，このような手法の適用を進めることに期待が高まっている．

本論文では，前述の課題を踏まえ，RNA-タンパク質複合体予測へのデータ駆動型の取り組みとして，RNA-タンパク質複合体構造の妥当性を評価するモデルの構築，および，XAI 技術を用いた選択結果の説明性の向上について検討した結果について報告する．この複合体構造妥当性評価モデルの構築プロセス，および，複合体構造予測プロセスにおいて想定する妥当性評価モデルの利用について**図 1** に示した．

複合体の 3 次元構造を入力とする深層学習モデルとしては，3 次元畳み込みニューラルネットワーク（3D-CNN: 3-dimensional convolutional neural network）を用い，RNA-タンパク質の組み合わせに対して推定される複合体構造候補の妥当性を評価するモデルの作成を行った．3D-CNN では，複合体構造を 3 次元空間での正規格子単位とする体積要素であるボクセル（Voxel）として表現し，タンパク質のアミノ酸種，RNA の核酸種が各ボクセルにチャネルとしてマップされる．作成された複合体構造の妥当性評価モデルに対して，XAI 技術の一つである Integrated Gradients[5]を用いて判定に重要となるボクセルを抽出し，複合体形成に重要と考えられるタンパク質-RNA の表面間の距離との関係について解析を行った．

以降の 2 章にて本報告で検討を行った方法について説明し，3 章にて結果と考察を提示したのち，結びとして 4 章にて今後の展望について述べる．

2 方法

本章では，RNA-タンパク質複合体構造の妥当性評価モデルについて，学習データ，評価モデルの作成，適用した XAI 技術について述べる．

2.1 学習データ収集と処理

複合体構造の妥当性判定モデルの学習データは，X 線結晶構造解析に基づいて決定された RNA-タンパク質複合体の 3 次元構造データ（PDB ファイル形式）である．実際のデータは，PDBbind データベース[6]から取得した．得られたデータセットから，大きな RNA 分子（水素を除く重原子数が 3,000 原子を超えるもの）を除外し，244 構造を抽出した．さらに，複数の類似した複合体構造が存在することによるデータリークを避けるため，RNA 配列およびタンパク質配列各々でローカルアライメントを実行し，RNA 配列とタンパク質配列の類似度がともに 40%以上となる類似複合体を特定して除外した．その結果，149 構造のユニークな複合体が選択された．次いで，生体分子向けに用いられる分子動力学ソフトウェア AMBER（Assisted Model Building with Energy Refinement）[7]を用いて，複合体構造の各々に対してエネルギー最小化を行った．

シミュレーションなどで推定された複合体構造の妥当性判定のための正解／不正解となる教師データの作成のため，エネルギー最小化した複合体構造から，RNA とタンパク質を分離して，再ドッキングを行った．再ドッキングには，タンパク質同士のドッキングソフトウェア ZDOCK[9]を，RNA-タンパク質のド

ッキング用に特別に調整したソフトウェア [10]を用い，各複合体に対して 2,000 通りのドッキング構造データを作成した．これらのドッキング構造から互いに異なる多様な構造を選択するため，タンパク質の原子位置を一致させたうえで，RNA の原子位置に基づいたクラスタリングを行った．クラスタリングには，k-近傍アルゴリズムを用い，2,000 通りの構造から 11 のクラスタに分割して，各々のクラスタからセントロイドを選び代表的な 11 構造を選択した．代表的な 11 構造とエネルギー最小化した複合体構造の類似度を，対応する原子位置の RMSD（Root Mean Squared Displacement）を算出し，RMSD が 5Å 以下の構造を自然な構造, 10Å 以上の構造を不自然な構造とした．自然な構造とエネルギー最小化した構造を合わせて正解データ（1），不自然な構造を不正解データ（0）として, 教師データとしてのラベル付けを行った．作成した学習データセットには, 計 1,734 構造が含まれる．

3D 畳み込みニューラルネットワーク（3D-CNN）の入力用のボクセルデータについては，これらの 1,734 構造の各々について，以下のようにデータ構造と値の割り付けを行って作成した．データ構造については，ボクセルサイズ 50×50×50 の 3 次元画像となるが，各ボクセルには，さらに 20 種類のアミノ酸と 4 種類の核酸を表現する計 24 チャネルを設定した．

実際の 3 次元空間との対応としては，x，y，z 方向の幅を 50Å とする立方体領域を用意して 3 次元画像に対応させ，これを隙間なく充填する x，y，z 方向の幅 1Å の立方体の体積要素をボクセルに対応させる．

各ボクセル・各チャネルへの値の割り付けでは，まず，タンパク質と RNA の複合体形成に重要となる相互作用領域がボックス内に確実に含まれるよう，タンパク質と RNA との最近接原子間の中点が立方体領域の中心に位置するように，複合体構造全体を平行移動させた．次いで，各々の残基／核酸について含まれる原子の重心座標を算出，その重心位置を中心とする 3 次元ガウス関数が各ボクセルの中心位置で持つ関数値を残基／核酸種に応じたチャネルに割り付けた（複数の残基／核酸から重複する寄与があれば加算）．以上にて，1,734 構造×24 チャネル×50×50×50 の配列データが作成された（ファイルサイズは，バイナリ形式で約 44GB）．データ拡張として，設定したデータ構造の下では，立方体領域を同一とするような 24 通りの回転操作による拡張が可能であり，これにより複合体構造数は，41,616 構造となる．データ容量の削減のため，これらの回転操作は，ミニバッチを抽出する際に行うこととした．

以上の操作によって作成されたボクセルデータの例を**図 2** に示す．

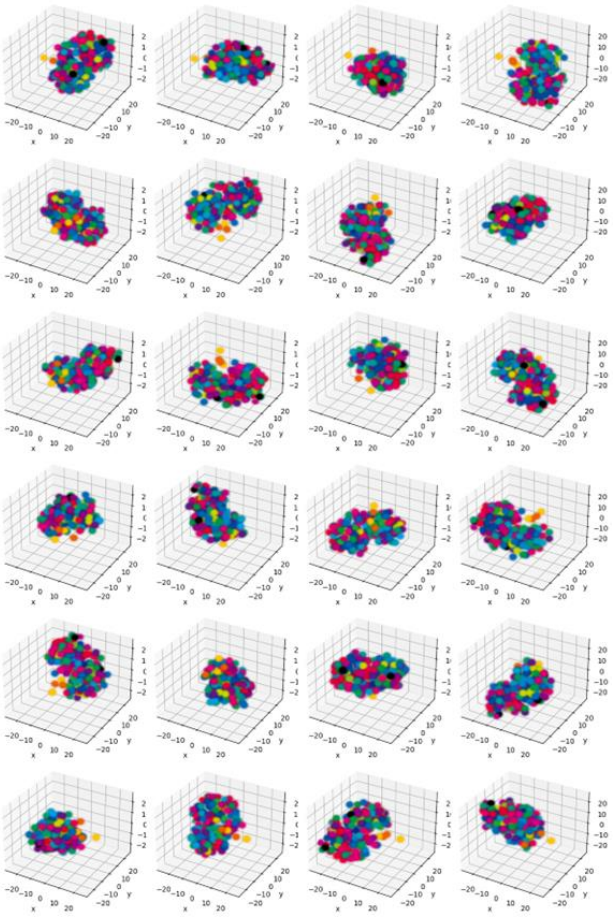

図 2　RNA-タンパク質複合体の 3 次元構造のボクセルデータ化： PDBID 6GX6 をもとにした単一の複合体構造に対して 24 通りの回転を行った構造．黄色～橙色が RNA の塩基，その他はタンパク質の残基に対応．

2.2　妥当性評価モデルの作成

複合体の妥当性判定モデルは，2.1 節にて説明した 24 チャネル×50×50×50 のデータを入力とし，妥当性判定指標となる単一の値を出力する関数である．

3 次元畳み込みニューラルネットワークを用いたたモデルの構造については，低分子とタンパク質のドッキング構造評価を目的とした先行研究 [11]を参考として検討を行った．

本報告で採用したニューラルネットワークモデルの構造を**図 3** に示す．モデルには 3 層の 3 次元畳み込み層を含み，各層の後に ReLU や ELU などの活性

化関数を設定した．また，2 番目，3 番目の 3 次元畳み込み層の後に，それぞれ，3 次元の平均プーリング層と最大プーリング層を設定した．これらの操作の後に，配列要素をフラット化して線形結合を行う全結合層および Softmax 関数により[0,1]の値を出力するものとした．

モデルの訓練では，全データを正解データおよび不正解データの数が均等となるよう 5 分割して用いた．5 分割したデータから 3 セットを選んで訓練データ，1 セットを検証データ，残りの 1 セットをテストデータとして用いて，モデルの訓練とテストを行った．この訓練・テストについて，5 セットの各々がテストデータとなるように，残りのセットから訓練データ，検証データを割り当てて計 5 回実施し，5 回の結果を総合してモデルを評価した．訓練時には，検証データの再現誤差を指標として学習の進行をモニタリングし，最良と推測されるモデルを選択した．得られたモデルに対して，テストデータを用いて AUC（Area Under the Receiver Operating Characteristic Curve）を算出して予測精度の評価指標とした．AUC は，二値分類問題で閾値を事前に定義することが困難な場合に，優れた性能評価指標となっている．

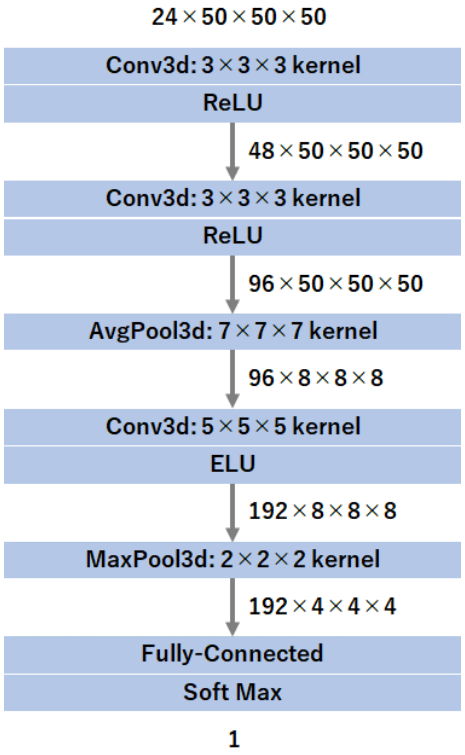

図 3　3D-CNN を用いたニューラルネットワーク構造（3D-CNN モデル）とデータサイズ

2.3　説明可能な AI 技術の適用

説明可能な AI（XAI）技術は近年注目を集めており，画像，テキスト，その他のデータタイプに対する機械学習モデルの予測の根拠を評価することが可能になってきている．本報告での複合体の妥当性判定モデルの入力はボクセルデータ（3 次元画像）であるため，通常（2 次元）の画像データ解析において多くの実績が報告されている Integrated Gradients（IG）法を用いることとした．IG 法では，予測結果に対する特徴量の寄与を評価する手法が満たすべき公理として，感度，実装不変性，完全性などを充足していることが示されている [5]．感度に係る公理では，モデルの予測に影響を与えた特徴量を正しく重要と判断することを保証しており，一般的な勾配（Gradient）ベースでの手法で勾配が 0 となる領域での問題を克服している．実装不変性は，モデルの詳細な実装が異なっていても同一の入力に対して同一の出力を返すものであれば，特徴量の寄与度は常に同一であることを保証する．最後の完全性は，モデルの予測値の増減が，どの特徴量によってどの程度引き起こされたのかを完全に分解して説明できることを保証する．これらによって信頼性が高く，説明力の高い特徴量の寄与度を提示する手法となっている．

3　結果と考察

本章では，複合体構造の妥当性評価モデルの学習，性能，および，説明可能な AI の適用の結果について述べる．

3.1　学習

3D-CNN を用いた複合体構造の妥当性評価モデル（**図 3**）の学習では検証誤差の最小値をモニタリングし，直前の 400 エポックの間検証誤差の最小値が更新されない場合に学習を停止させることとした．訓練時の学習曲線（5 セット分）を**図 4** に示す．訓練の開始から 400 エポックを過ぎるころから訓練誤差，検証誤差ともに減少傾向となり学習が進んだ．その後，750～1,000 エポックで検証誤差は最小となり，増加傾向に転じ，過学習の状態となった．これらは典型的な学習の進展であり，適切に学習が進んだことを示している．各セットで検証誤差が最小になった時点のパラメータを用いて，以降の予測性能の検証を行った．

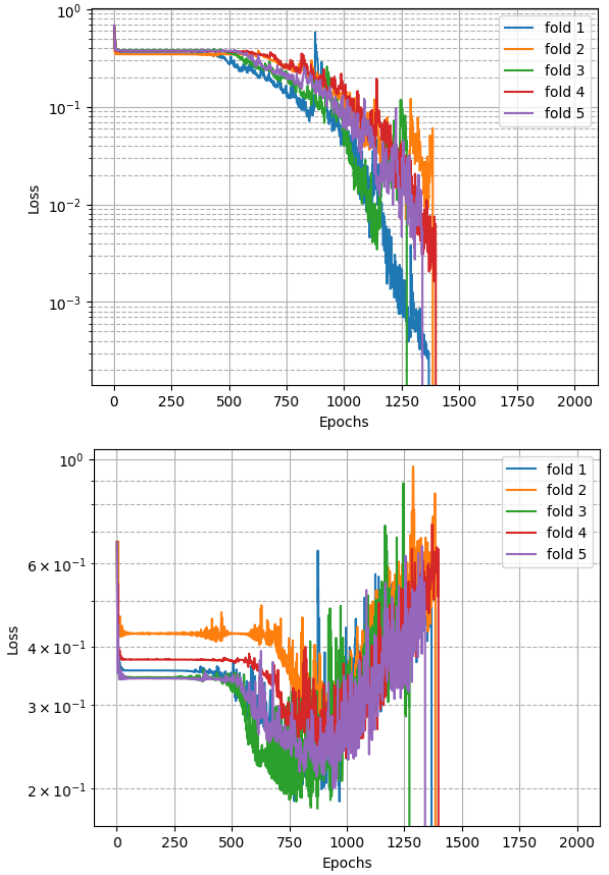

図 4　ニューラルネットワークモデルの学習曲線（テストデータ 5 セット分，上：訓練誤差，下：検証誤差）

3.2　性能検証

複合体の妥当性を評価する 3D-CNN モデルの性能については，訓練データセットに教師データとして付与した正解/不正解のラベル（1/0），および，3D-CNN モデルの出力である妥当性の指標[0,1]に基づいて算出した ROC AUC によって行った．

ベンチマークとして，RNA-タンパク質複合体構造予測シミュレータである，3dRPC[12,13]，DARS-RNP および QUASI-RNP[14,15]を用いた．これらにより，RNA-タンパク質の複合体構造データ（PDB ファイル形式）を入力として，RNA とタンパク質の相互作用エネルギーを算出した．一般的に，相互作用エネルギーの低い状態が，複合体として妥当であると解されるため，算出した相互作用エネルギーを複合体構造の妥当性の指標とし，3D-CNN モデルと同様に正解/不正解のラベル（1/0）と合わせて，AUC を算出した．

3D-CNN モデルと，ベンチマークのシミュレータの性能を ROC（Receiver Operating Characteristic）曲線および AUC で比較した結果を**図 5** に示す．3D-CNN は，他のシミュレータと比較して同等の性能を示している．ベンチマークとして用いたシミュレータは，本報での複合体構造の妥当性評価という目的とは異なるが，RNA-タンパク質複合体の X 線結晶構造解析データを再現するように物理・化学的原理に基づいて構築・改良がなされてきたものである．一方，3D-CNN による複合体構造の妥当性評価モデルは，シミュレータと比較して単純なデータ（基本的に 3 次元構造情報と核酸塩基種・アミノ酸残基種）のみに基づいてモデルが構築されているが，遜色のない性能を実現できている．

シミュレータで生成される多数の構造候補から“正解”を選ぶ際に困難が生じている現実に反して，シミュレータによる妥当性の評価は想定より高いレベルにあるといえる．これは，今回のテストデータには，X 線結晶構造解析構造に対する RMSD が 5Å～10Å となるような中間領域の複合体構造が含まれていないことに起因していると推測される．このことはさらに，シミュレータによる構造の妥当性判断に用いたエネルギー指標は，大域的な構造の違いを区別する場合には明確であるが，より詳細な構造の違いに対しては曖昧になることを示唆しているともいえる．

本報告で参考にした小分子・タンパク質のドッキング構造の妥当性評価スコアの検討 [11]では，本報告とは対照的に，妥当性評価モデルの方がシミュレータを上回る性能を示すことが報告されている．同報告によると，訓練・検証データには本報告と同様に中間領域を除いているが，テストデータには中間領域の構造を含めており，前述の推測をサポートする．この推測を確認するため，本報告の取り組みにおいても中間領域のデータを含めた追加検証が必要である．

今後，RNA-タンパク質複合体構造についての X 線結晶構造解析のデータが蓄積されるにつれて，データに基づく複合体構造の妥当性評価のさらなる性能向上も期待できる．これを踏まえると，構造未知の複合体に対する複合体構造予測を行う場合に，シミュレータにより生成される多数の複合体構造候補の中から，より自然な構造を絞り込むというタスクの中で，妥当性評価モデルの活用による効率の良い作業が実現できるものと期待される．

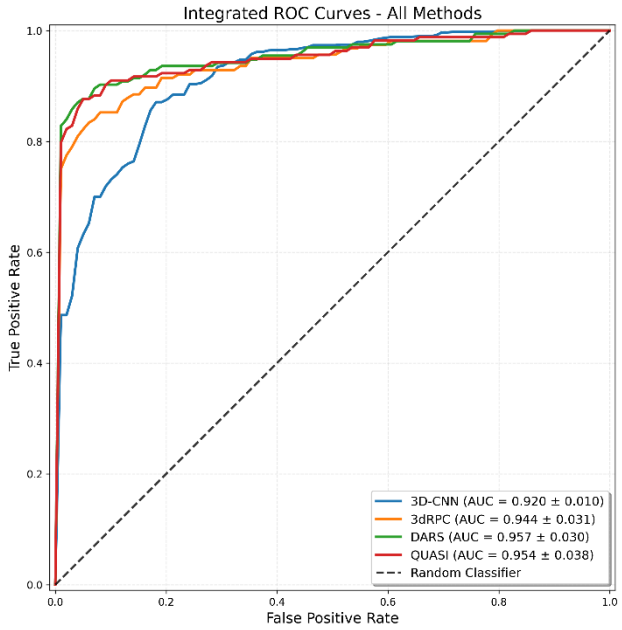

図 5　3D-CNN とシミュレータ（3dRPC, DARS-RNP, QUASI-RNP）による複合体構造妥当性評価の性能比較（ROC 曲線および AUC）

3.3　説明性

3D-CNN による複合体構造の妥当性評価の根拠を分析するための XAI 技術として IG 法を採用して検討を行った.

IG 法を用いると, RNA-タンパク質の複合体構造のボクセルデータに対して，複合体構造の妥当性評価値への寄与の大きさ各ボクセルに割り付けることができる. さらに, この寄与を各ボクセル位置する核酸塩基／アミノ酸残基各々からの寄与として集約することで，複合体の妥当性評価の根拠を物理的な実体と関連付けて示すことができる.

図 6 には，構造の大きさが異なる 2 つの事例について，IG 法から算出した核酸塩基／アミノ酸残基寄与の大きさに基づいて強調表示した複合体構造の模式図を示す. 相互の距離（最近接原子間距離）が短い核酸塩基とアミノ酸残基間を強調表示した模式図を併せて図示した.

IG 法と距離に基づく強調表示領域は類似しており, RNA とタンパク質の界面に近い箇所で IG 法による妥当性の評価値が高くなっていることが見られる. 一方で，両者は完全に一致しているわけではない.

距離が一定の範囲にある核酸塩基・アミノ酸残基のリストと，距離に基づくリストと同数の核酸塩基・アミノ酸残基を IG への寄与の大きいものから順に選択したリストの一致率を，ランダムに選択した 40 構造にわたり算出したときの一致率のヒストグラムを**図 7** に示す. 距離が 5Å の範囲内では，平均して 50%程度の一致率であり，核酸種およびアミノ酸残基種に基づく相互作用の強さの違いが表現されているものと推測される. この仮説については，例えば，量子力学に基づくフラグメント分子軌道法 [16]など，核酸種およびアミノ酸残基種間の相互作用を詳細に分析する手法を適用することにより，明らかにできると考えられる.

さらなる詳細な解析は必要であるが，XAI 技術により，妥当性評価への各ボクセルの寄与を，物理・化学的な実体である核酸塩基・アミノ酸残基の情報と結びつけることで，RNA-タンパク質の複合体構造についての説明性を向上させ，構造妥当性判断において物理・化学的洞察を提示することが可能であることが示された.

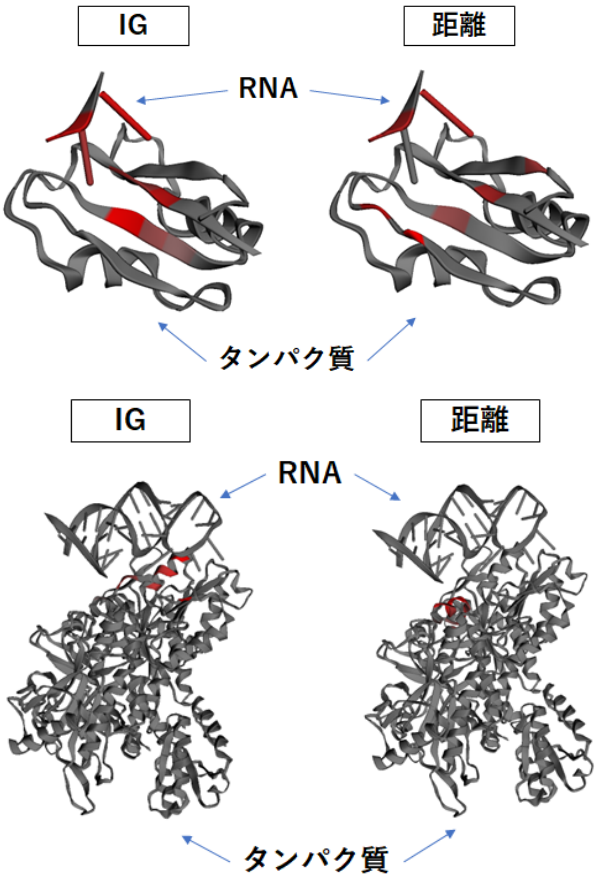

図 6　複合体構造（PDBID 上段：2L41，下段：1KOG）に対する IG 法による解析（左）と RNA-タンパク質間界面（右）：IG 法では複合体としての妥当性評価への寄与を寄与度に応じて赤で強調，界面については距離の近い核酸塩基-アミノ酸残基間を近さに応じて赤で強調

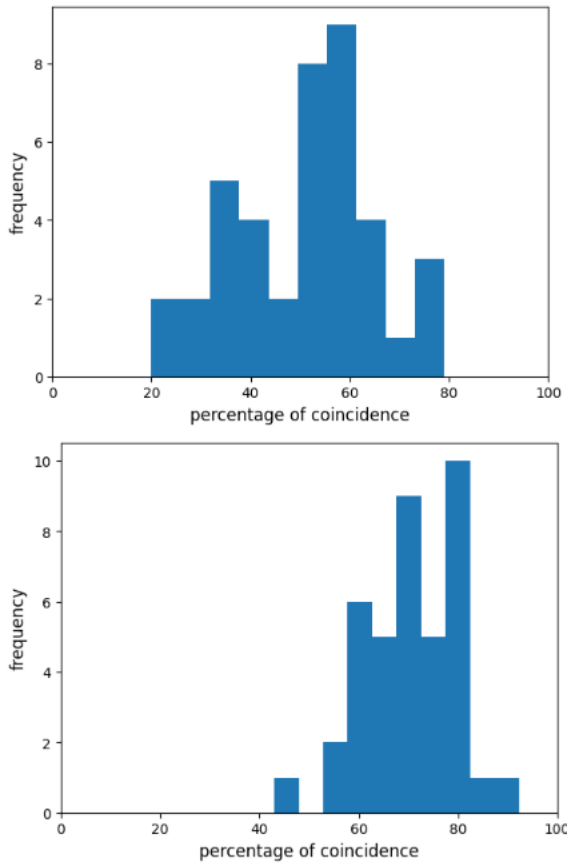

図 7　一定の距離範囲（上:5 Å，下 10 Å）にある核酸残基・アミノ酸残基リストと，IG 法による複合体妥当性判定への寄与が高い核酸塩基・アミノ酸残基リストの一致率（距離に基づき同数で比較）：ランダムに選択した 40 構造にわたり算出

4　おわりに

本報告では，RNA-タンパク質の複合体構造予測において，多数の複合体構造候補から妥当な構造を選択するためのデータ駆動型アプローチによる評価法の構築と説明性の向上を目的とした検討を行った．

複合体の 3 次元構造をボクセルデータで表現し，ボクセルデータを入力として構造の妥当性を判定するモデルを 3D-CNN をベースとして作成した．与えられた複合体構造候補に対する妥当性の判定能力は既存の複合体予測シミュレータと同レベルであり，今後の X 線結晶構造解析データの増加に伴う判定能力の向上が期待できる．また XAI 技術として，IG 法を用いることにより，構造妥当性への核酸塩基・アミノ酸残基の寄与を数値化・可視化することで，妥当性判定での物理・化学的洞察の提示が可能であることを示した．

今後，3D-CNN モデルについては，複合体構造の妥当性評価におけるシミュレータとの性質の差異，および，IG 法による構造妥当性への核酸塩基・アミノ酸残基の寄与と物理・化学的な解釈についての詳細解析を進めることで，複合体構造予測にて多数生成される構造候補からより実際に生じている可能性の高い構造を，高精度・高速かつ説明力を伴って絞り込むことが可能になると思われる．

このような技術が実現することで，遺伝子発現制御，RNA スプライシング，翻訳，ウイルス複製などの基礎的な生物学研究の強力な駆動力となるだけでなく，RNA アプタマーの配列設計など RNA が係る創薬プロセスを大きく変革することが期待される．

引 用 文 献

1) RCSB PDB, https://www.rcsb.org, Berman, Helen M., et al.: "The protein data bank." *Nucleic acids research* 28.1 (2000): 235-242.
2) Jumper, John, et al.: "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.
3) Corso, Gabriele, et al.: "Diffdock: Diffusion steps, twists, and turns for molecular docking." *arXiv preprint* arXiv:2210.01776 (2022).
4) Ketata, Mohamed Amine, et al.: "Diffdock-pp: Rigid protein-protein docking with diffusion models." *arXiv preprint* arXiv:2304.03889 (2023).
5) Sundararajan, Mukund, Ankur Taly, and Qiqi Yan: "Axiomatic attribution for deep networks." *International conference on machine learning.* PMLR, 2017. 3319.
6) PDBbind, http://www.pdbbind.org.cn, (参照 2023-5-11)
7) Salomon‐Ferrer, Romelia, David A. Case, and Ross C. Walker: "An overview of the Amber biomolecular simulation package." *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3.2 (2013): 198-210.
8) The Amber Home Page, https://ambermd.org/
9) Chen, Rong, and Zhiping Weng: "Docking unbound proteins using shape complementarity, desolvation, and electrostatics." *Proteins: Structure, Function, and Bioinformatics* 47.3 (2002): 281-294.
10) Iwakiri, Junichi, et al.: "Improved accuracy in rna–protein rigid body docking by incorporating force field for molecular dynamics simulation into the scoring function." *Journal of chemical theory and computation*

12.9 (2016): 4688-4697.
11) Ragoza, Matthew, et al.: "Protein–ligand scoring with convolutional neural networks." *Journal of chemical information and modeling* 57.4 (2017): 942-957.
12) Huang, Yangyu, Haotian Li, and Yi Xiao: "3dRPC: a web server for 3D RNA–protein structure prediction." *Bioinformatic*s 34.7 (2018): 1238-1240.
13) 3dRPC Web Server, http://biophy.hust.edu.cn/new/3dRPC
14) Tuszynska, Irina, and Janusz M. Bujnicki: "DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking." *BMC bioinformatics* 12.1 (2011): 348.
15) Laboratory of Bioinformatics and Protein Engineering, Statistical potentials for RNA-Protein docking, https://genesilico.pl/software/stand-alone/statistical-potentials
16) Kitaura, Kazuo, et al.: "Fragment molecular orbital method: an approximate computational method for large molecules." *Chemical Physics Letters* 313.3-4 (1999): 701-706.